prof. dr hab. Krzysztof Konarzewski
Instytut Psychologii PAN w Warszawie

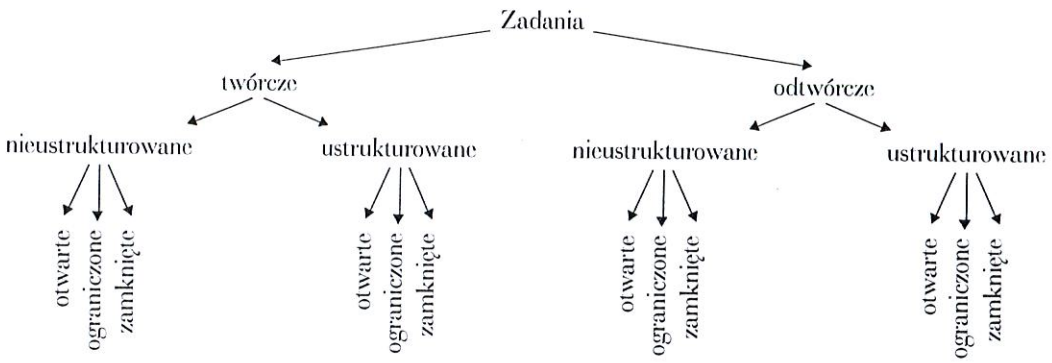
Miejsce testów wyboru w kulturze oświatowej

STRESZCZENIE. Test to ujednolicona procedura pobierania próbki określonego zachowania i jej interpretacji ilościowej. Na swoistość testu wyboru składa się nie tylko zamknięta forma odpowiadania, ale także — jak wykazuje analiza istniejących testów — odtwórczość i ustrukturowanie. Taniość testów wyboru sprzyja ich umasowieniu, a masowość — rozszerzaniu zastosowań. Użycie testów wyboru do rozliczania szkół umożliwia im ingerowanie w proces kształcenia, mimo jawnych wad i ateoretyczności,. Są dowody, że test wyboru mierzy co innego niż inne testy rzekomo tej samej wiedzy. Wybierając z pobudek ekonomicznych test wyboru, bezwiednie zmieniamy pojęcie wykształcenia i zmuszamy szkoły, by dawały uczniom coś, czego być może nie uznalibyśmy za wartościowe, gdybyśmy o tym pomyśleli na chłodno. Test wyboru należy do kultury taśmy produkcyjnej. W aprobacie tego testu można widzieć pochwałę socjalizacji, która wytwarza sprawnych wykonawców. Miejscem tej socjalizacji jest oświata dualna: jedna dla mas, którym wystarczą „kompetencje", inna dla elity, której będą potrzebne zdolności koordynowania wysiłku mas.

Czym jest test wyboru?

Test to ujednolicona (standardowa) procedura pobierania próbki określonego zachowania się i przyporządkowywania tej próbce liczby (wyniku surowego). Taka procedura składa się z instrukcji, zbioru zadań i wzorów ich wykonania. Instrukcja i zadania standaryzują warunki testowania, a wzory wykonania standaryzują wynik. Dzięki temu można porównywać dowolne dwa wyniki (z dokładnością wyznaczoną przez rzetelność testu), a dalej — znajdować rozkłady wyników w populacjach i przekształcać je w wygodniejszą postać (centylową, stenową itp.). Kiedy mówimy o standaryzowanych testach osiągnięć szkolnych, mamy na myśli zarówno standaryzację warunków i wyników surowych, jak i możliwość odniesienia tych wyników do populacji.

Na czym polega „różnica gatunkowa" testu wielokrotnego wyboru *(multiple choice test)*? Na pozór odpowiedź jest prosta — na tym, że do każdego zadania dołącza się kilka gotowych odpowiedzi i wymaga od badanego, by wybrał spośród nich właściwą. Próbka zachowania, której dostarcza test, to zatem ciąg wybranych opcji. Dopełnieniem testu wyboru byłyby: test krótkich odpowiedzi *(short answer)*, test swobodnych odpowiedzi *(free response)* itp. Można jednak wątpić o użyteczności klasyfikacji opartej na drugorzędnej cesze formalnej. Spróbujmy sklasyfikować testy osiągnięć w bardziej systematyczny sposób. Będą nam do tego potrzebne trzy względy podziału.



Pierwszy: *dominujący typ operacji umysłowych, których wymaga zadanie.* Zadania odtwórcze wymagają wyszukania w pamięci długotrwałej wiadomości lub procedur spełniających warunki zadania. Zadania twórcze wymagają pomysłu, czyli nowej konstrukcji umysłowej (syntezy warunków zadania i zasobów pamięci długotrwałej), spełniającego warunki zadania. To rozróżnienie może zawieść w odniesieniu do pojedynczego ucznia (zadanie twórcze może się okazać odtwórcze, jeśli uczeń pracował nad nim w przeszłości, a zadanie odtwórcze może się okazać twórcze, jeśli uczeń nie zapamiętał wymaganych wiadomości lub procedur), toteż trzeba je odnosić do programu kształcenia i objętej nim populacji uczniów.

Drugi: *liczba wskazówek ukierunkowujących wykonanie.* Zadania ustrukturowane zawierają dużo wskazówek: podają wszystkie wiadomości i procedury potrzebne do wykonania (lub jawnie odsyłają do nich) i gwarantują istnienie poprawnego rozwiązania. Zadania nieustrukturowane podają mało wskazówek albo topią je w morzu informacji nieistotnych lub mylących. Nie sugerują etapów pracy i nie dają ścisłych kryteriów odróżniających poprawne wykonanie od niepoprawnych.

Trzeci: *forma udzielania odpowiedzi.* Może być otwarta *(free response)*, ograniczona *(short answers)* lub zamknięta *(multiple choice)*.

Zdawałoby się, że trzeci wzgląd podziału jest niezależny od dwóch poprzednich, ale od dawna wiadomo, że zamknięta forma odpowiadania znacznie częściej towarzyszy zadaniom odtwórczym i ustrukturowanym niż jakimkolwiek innym. W będących w obiegu testach wyboru można czasem znaleźć zadania twórcze, np.:

> «Nasz car ma wspaniałe plany: zdobycia Mandżurii i przyłączenia Korei. Marzy także o zostaniu panem Tybetu. Pragnie opanować Persję i zdobyć nie tylko Bosfor, lecz także Dardanele». Z którego roku pochodzi ta wypowiedź? 1775, 1815, 1835, 1903, 1913 (SAT, 34% poprawnych odpowiedzi).

> Dla każdej dodatniej liczby całkowitej n zdefiniujmy $^{\Delta}n$ jako liczbę otrzymaną z przepisania cyfr liczby n w odwrotnej kolejności i odrzucenia początkowych zer, np. $^{\Delta}5 = 5$, $^{\Delta}30 = 3$, $^{\Delta}125 = 521$. Które z poniższych twierdzeń muszą być prawdziwe dla wszystkich dodatnich liczb całkowitych:
> I. $^{\Delta}(^{\Delta}n) = n$; II. $^{\Delta}(10n) < 10n$; III. $^{\Delta}(1 + n) = 1 + {}^{\Delta}n$. *Żadne; Tylko I; Tylko II; Tylko I i II; Wszystkie*
> (SAT, 34% poprawnych odpowiedzi)

> Właściwe wykonanie trudnego zadania wymaga od ucznia silniejszej motywacji niż wykonanie łatwego zadania. Prawda; Fałsz (40% poprawnych odpowiedzi)

Ale takie zadania to rodzynki w masie zadań typu „Jak nazywamy...?", „Kto...?", „Ile...?" itp. Nawet testy opracowane przez renomowane ośrodki mają co najmniej połowę pozycji wymagających prostej reprodukcji wiadomości i formuł. Levine i in (1970) opisali wysiłki ulepszenia testu mającego mierzyć złożone umiejętności medyczne. W pierwszej wersji — jak jednomyślnie uznali eksperci — ponad połowa zadań wymagała jedynie odtworzenia wiadomości, a mniej niż czwarta część — interpretacji, oceny lub zastosowania twierdzenia. Podczas prac nad drugą wersją dostarczono autorom testu wyraźnych wskazówek, jak budować twórcze zadania, i zapewniono pomoc recenzentów. Mimo to większość zadań nowej wersji nadal mierzyła zapamiętanie wiadomości. Testy wielokrotnego wyboru okazują się zatem testami odtwórczymi, ustrukturowanymi i zamkniętymi.

Powody upowszechniania się testów wyboru

Na dużą skalę test wyboru (Army Alpha) zastosowano po raz pierwszy w armii amerykańskiej podczas I wojny światowej. Potem zainteresowało się nim szkolnictwo, zwłaszcza wyższe, borykające się z problemem selekcji kandydatów. W 1926 r. pojawił się pierwszy SAT (Scholastic Aptitude Test), którym objęto 8 tys. uczniów. W latach 80. liczbę pomiarów testowych szacowano na 30 mln rocznie (Frederiksen 1984). Dziś jest ich zapewne więcej.

U podłoża tej oszałamiającej kariery leży fakt, że test wyboru daje próbkę zachowania się, którą można tanio przekształcać w wynik liczbowy. Tanio — ponieważ mechanicznie i szybko. Nie trzeba zatrudniać ani licznego, ani wykwalifikowanego personelu. Niskie koszty jeszcze bardziej zmniejszyły elektroniczne czytniki, które pojawiły się po II wojnie. W 1984 r. agencja ETS *(Education Testing Service)* dysponowała maszynami, które przetwarzały 10 tys. arkuszy odpowiedzi na godzinę. Dziesięć lat później maszyny przetwarzały 15 tys. arkuszy na godzinę.

Taniość sprzyja umasowieniu, a masowość — rozszerzaniu pola zastosowań. Od końca lat 70. wykorzystuje się wyniki testowania do podejmowania decyzji nie tylko o uczniach, lecz także o szkołach, w ramach accountability: „rozliczania" szkoły przez bliższe i dalsze otoczenie społeczne. Wyniki testowania uczniów stały się miarą jakości usług oświatowych. Udostępniano je obwodom, szkołom i nauczycielom, publikowano w prasie. Szkoły, których uczniowie wykonywali testy poniżej normy, były poddawane inspekcji. To wywierało stały nacisk na nauczycieli, by uczyli tego, czego wymagają testy. Mnożyły się szkolenia, na których rozkładano zadania testowe na szczegółowe umiejętności i instruowano nauczycieli, jak ich nauczać. Udostępniano też stare testy „dla celów edukacyjnych".

Ta polityka szybko przyniosła owoce. W latach 70. wyniki testowania pogarszały się z roku na rok. W latach 80. zaczęły rosnąć. Badania psychologiczne popierały doniesienia statystyki oświatowej. Badacze stwierdzali, że częste wykonywanie testów polepsza zapamiętanie wiadomości i podwyższa wyniki testowania. Uczniowie, którzy uczyli się z myślą o teście, wykonywali go lepiej (zwłaszcza pozycje wymagające krótkiej odpowiedzi) niż uczniowie, którzy nie oczekiwali testu i uczyli się po to, „by się nauczyć". Dużą popularnością cieszyły się porównania dydaktycznej skuteczności różnych „formatów" testu. Stwierdzono np., że test krótkich odpowiedzi bardziej sprzyjał zapamiętaniu informacji niż test wyboru i że przyrost wykonania w serii testów wyboru był mniejszy niż przyrost wykonania w serii testów uzupełniania, co mogło wynikać stąd, że obcowanie z opcjami niepoprawnymi interferuje z pamiętaniem opcji poprawnych.

Zauważmy: za sprawą techniki testy, które zrazu miały orzekać, czy uczeń opanował treści nauczania, upowszechniają się i zaczynają wpływać na same te treści i proces nauczania. Dzięki temu wyniki testowania polepszają się, testy rosną w prestiż, a lobby ich producentów żyje dostatniej. Kraje, w których dotąd stosowano „intuicyjne", interakcyjne ocenianie osiągnięć szkolnych, gwałtownie modernizują swoją oświatę, powołując organizacje do spraw egzaminów testowych. A przecież testy wyboru mają liczne wady — jawne i ukryte.

Jawne wady testów wyboru

Wrażliwość na błędy testu. Zacznijmy od oczywistego faktu: instrukcje, zadania i opcje odpowiedzi są usiane błędami — merytorycznymi, językowymi i typograficznymi. Nie widziałem bezbłędnego testu (nie wyłączając tych, które sam budowałem). Zapoznając się z testami Instytutu Badania Kompetencji, natknąłem się na zadanie „Zapisz za pomocą wyrażenia algebraicznego liczbę trzy większą od x". Mój umysł, chcąc nadać sens temu zdaniu, automatycznie wstawił „razy", więc odpowiedziałem 3x. Niestety, umysł autorów działał inaczej, bo wzór dopuszczał tylko: $x+3$; $3>x$; $x<3$.

A oto przykład bardziej subtelnego błędu. W SAT II jest takie zadanie:

> Cztery boki ośmiokąta foremnego leżą na bokach kwadratu. Długość boku kwadratu wynosi 1. Jaka jest długość boku ośmiokąta? 1/3; $1/(1+\sqrt{2})$; 1/2; $\sqrt{2}/3$; 1/8.

W wywiadzie po wykonaniu testu jeden z lepszych uczniów klasy VIII tak tłumaczył swoją porażkę w tym zadaniu:

> Obliczyłem, że bok równa się $\sqrt{2}/(2+\sqrt{2})$, ale nic takiego nie było w odpowiedziach. Pomyślałem, że trzeba to doprowadzić do lepszej postaci, więc pomnożyłem licznik i mianownik przez $(2-\sqrt{2})$ i wyszło $\sqrt{2}-1$. Tego też nie było. Rozwiązałem zadanie jeszcze raz, ale wyszło tak samo. Wtedy zacząłem się trochę denerwować, bo czas uciekał. Pomyślałem, że wrócę do tego zadania później...

W istocie oba rozwiązania są równoważne drugiej opcji.

Wrażliwość na niestaranne czytanie. Zauważmy, że skrajne etapy testowania: zadanie i odpowiedź są widoczne, ale tego, co pośredniczy między nimi, nie widać. Domyślamy się, że uczeń musi zrozumieć zadanie i sięgnąć do zasobów pamięci semantycznej i proceduralnej. Ale jest wiele sposobów zrozumienia, wiele strategii dostępu do zasobów i wiele metod ich wykorzystania. Uczeń, który zakreśla złą opcję, może wiedzieć, a uczeń, który zakreśla dobrą — nie wiedzieć tego, co jest przedmiotem pomiaru. Dlaczego tak bywa?

Szczegółowa analiza odpowiedzi często prowadzi do wniosku, że uczniowie nieuważnie czytają tekst zadania. Trzeba np. podać największą liczbę całkowitą spełniającą nierówność $x<-3$. Ok. 30% uczniów zakreśla –2, a 17% –3,1. Pierwsi nie rozumieją pojęcia spełniania nierówności, a drudzy rozumieją (ich błąd wynika z przeoczenia warunku „całkowita"). Dwie błędne odpowiedzi nie są więc równoważne. Podobnie z podchwytliwym zadaniem w teście wiedzy religijnej. Uczeń miał orzec, czy prawdą jest, że „Jezus Chrystus jest człowiekiem, który stał się Synem Bożym dla naszego zbawienia". Zadanie to okazało się bardzo trudne (odpowiedzi „Nie" udzieliło jedynie 6% uczniów) i całkowicie nietrafne (współczynnik korelacji z łącznym wynikiem był zerowy). Zadanie mierzyło raczej staranność czytania niż znajomość katechizmu. Tylko patrzeć, jak zaczniemy uczyć dzieci czytania instrukcji i zadań testowych ze starannością wymaganą dotąd tylko od sapera.

Wrażliwość na zgadywanie. Jeśli wiem — odpowiadam, jeśli nie wiem — zgaduję. Dostępność tej strategii wykonywania testu wyboru od dawna spędza sen z oczu konstruktorów. Jedni radzą umieścić próg zaliczenia powyżej wyniku, który można osiągnąć metodą zgadywania (w przypadku testu „prawda—fałsz" powyżej połowy zadań), godząc się, by drugorzędna cecha testu kontrolowała ważną decyzję oświatową. Inni wymyślają różne poprawki na zgadywanie, choć krzywdzą one tych, którzy nie zgadywali. Dobrego wyjścia nie widać.

Ateoretyczność. Zauważmy, że testy osiągnięć szkolnych to wynalazek czysto empiryczny. Widzimy, że działają, ale nie wiemy, dlaczego działają i kiedy działają nie tak, jak oczekujemy, ponieważ nie dysponujemy teorią psychologiczną, która wiązałaby wykonanie testu ze zmieniającymi się zasobami indywidualnego umysłu.

Weźmy np. zdroworozsądkowe założenie, że wyniki testowania rosną w miarę rozwoju poznawczego ucznia. Przyjmuje się je np. w Anglii i Walii. W *National Curriculum* każdy standard osiągnięć *(attainment target)* jest rozłożony na 10 poziomów osiągnięć *(levels of achievement)*, które przyporządkowuje się liniowo wiekowi ucznia (np. siedmiolatek powinien znajdować się na poziomie od 1 do 3, a szesnastolatek na poziomie od 4 do 9). Czy jednak nie bywa tak, że zrozumienie pojęcia (zwłaszcza niefenomenalistycznego lub kontrintuicyjnego) prowadzi przejściowo do obniżenia wykonania testu? Czy źródłem błędów nie może być to, że dziecko zaczyna uświadamiać sobie problematyczność zagadnienia, bierze pod uwagę więcej przesłanek, podejmuje wewnętrzny dialog z pytającym? Gdyby tak było, niektóre błędy świadczyłyby o postępie poznawczym. Nie znam wyników badań bezpośrednio potwierdzających taką możliwość, ale sam odkryłem kiedyś takie zadanie geometryczne, które dzieci z klasy zerowej wykonują znacznie lepiej niż uczniowie klasy IV (Konarzewski 1985). Być może ci drudzy zbyt szeroko wykorzystują nową zdobycz poznawczą (strukturę operacji konkretnych), być może ich wiedza przejściowo traci spójność. W podobnym duchu wypowiada się wybitny fizyk (Rogers 1986, s. 107): przeciętni uczniowie wpisują się w myślenie egzaminatora, wybitni mogą dostrzec, że inna opcja jest jeszcze lepsza. Na egzaminie ustnym mają szansę przekonać do niej komisję, ale z testem niepodobna dyskutować.

Inna zagadka: od dawna wiadomo, że w typowych testach osiągnięć lepiej wypadają ekstrawertycy niż introwertycy. Czy jest tak dlatego, że pierwsi lepiej się uczą, bo mają — jak się niemądrze powiada — większy zasób energii, czy może dlatego, że sprzyja im sytuacja testowania (ekstrawertycy mogą myśleć jednowymiarowo, ignorować uboczne skojarzenia)? Dlaczego lęk przed egzaminem (test anxiety) jednym pomaga, innym przeszkadza? Tego rodzaju fakty sugerują, że wynik testowania w jakiejś części zależy od procesów kontrolowania emocji. Niestety, nic o tym nie wiemy — ale nie przeszkadza to w budowaniu krajowego systemu egzaminów testowych.

Ukryte wady testów wyboru

Trafność testu wyboru

Czy testy wyboru mierzą to samo, co inne testy? — to pytanie ma podstawowe znaczenie. Ze strony badaczy najpierw padały uspokajające odpowiedzi, ponieważ wykrywali oni bardzo wysokie współczynniki

korelacji (w okolicy 0,9) między wynikami testów wyboru i testów krótkich odpowiedzi. Dokładniejsze analizy (konfirmacyjna analiza czynnikowa) ujawniały czasem drugorzędny czynnik formy odpowiadania, ale tylko w testach rozumienia tekstu, a nie w testach rozumowania matematycznego[1] (np. Traub i Fisher 1977). Niestety, te wyniki, krzepiące zwolenników testów wyboru, są uogólniane nieprawomocnie. Porównywanie testów różniących się formą odpowiedzi ma sens jedynie przy ustaleniu wyższych poziomów klasyfikacji. W znanych mi badaniach autorzy brali gotowy test wyboru i na jego podstawie budowali test krótkich odpowiedzi. Wniosek, że forma odpowiadania nie robi różnicy, trzeba więc ograniczyć do testów odtwórczych ustrukturowanych.

Jak jest w testach innego rodzaju, dowiadujemy się z pomysłowego badania Frederiksena i Warda (1978). Zbudowali oni test *(Formulating Hypotheses)*, który nazwałbym twórczym, ustrukturowanym i otwartym. Każdą pozycję testu stanowił krótki opis pewnego badania (z dziedziny nauk o zachowaniu) i jego wyników (w postaci tabeli lub wykresu oraz komentarza). Badany miał wysunąć możliwe wyjaśnienie wyników i sformułować hipotezy konkurencyjne, które należałoby rozważyć[2]. Na podstawie faktycznie udzielanych odpowiedzi autorzy stworzyli zamkniętą wersję tego testu. Współczynniki korelacji między wynikami wersji otwartej i zamkniętej okazały się bardzo niskie (jakość pomysłów: 0,18, liczba pomysłów: 0,19). Wniosek jest prosty: obie wersje nie mierzą tego samego konstruktu.

Jak się przekonać, co mierzy każda z nich? Autorzy włączyli do analiz wyniki typowych testów psychologicznych. Okazało się, że czynniki: wiadomości i rozumowanie słowne korelowały podobnie z wynikami obu wersji testu, ale czynnik Guilforda: płynność ideacyjna korelował tylko z wersją otwartą. Ponieważ płynność ideacyjna to z grubsza tyle, co szybkość przeszukiwania zasobów pamięci długotrwałej w celu wysunięcia odpowiedniego pomysłu, można uznać, że wykonanie testu w wersji zamkniętej nie wymaga generowania pomysłów. Znaczy to, że zmiana formy odpowiedzi znosi (lub ogranicza) twórczy charakter testu.

Podobnemu badaniu (Frederiksen 1984) poddano test twórczy, nieustrukturowany i otwarty, czyli zbiór zadań, które Simon (1973) nazwał był problemami słabo ustrukturowanymi *(ill-structured)*. Każda pozycja testu formułuje problem w sposób niepełny i nie zawiera wszystkich wiadomości potrzebnych do rozwiązania. Zadaniem badanego jest wymyślić hipotetyczne rozwiązanie, poprosić o dodatkowe informacje, sprawdzić w ich świetle swój pomysł, wysunąć następny — i tak do momentu, gdy będzie pewny rozwiązania. Na podstawie wykonania wersji oryginalnej sporządzono wersję zamkniętą i porównano wyniki. Analiza wykazała, że w wykonaniu wersji otwartej, zwłaszcza w odniesieniu do pomysłów, których bezpośrednio nie sugerują dostarczone informacje, bierze udział czynnik płynności ideacyjnej, a w odniesieniu do wnioskowania — czynnik rozumowania. W wersji zamkniętej wtórne czynniki wyników były słabe i nieskorelowane ani z płynnością, ani z rozumowaniem. Najwyraźniej studenci badani wersją zamkniętą najpierw zapoznawali się z instrukcją i opisem problemu, potem czytali pierwszą opcję odpowiedzi, przeszukiwali zasoby pamięci, by stwierdzić, czy jest poprawna, zaznaczali, jeśli taką się wydawała, i przechodzili do następnej opcji. Wersja zamknięta eliminowała generowanie pomysłów.

Do testów twórczych, nieustrukturowanych i otwartych można zaliczyć tzw. testy wykonania *(performance tests)*. Ta nazwa wydaje mi się wysoce niefortunna, ponieważ prowadzi do dziwacznych fraz w rodzaju „wykonanie testu wykonania"[3], a co ważniejsze, mylnie sugeruje, że główna linia podziału testów biegnie między poleceniem „powiedz/napisz" a „zrób". W testach inteligencji (np. WAIS-R) odpowiada temu podział na skale słowne i bezsłowne. Ale w testach osiągnięć „zrób" znaczy też „powiedz". Dlatego lepiej upatrywać różnicy w poziomie wskazówek (struktury). Testy wykonania *wzywają do stworzenia dzieła* (wygłoszenia miniwykładu, napisania eseju, przeprowadzenia dowodu matematycznego, zbadania chłonności papierowych ręczników trzech marek itp.). Inne testy *stawiają pytanie*, na które trzeba odpowiedzieć. Typowe pytanie

[1] Najprawdopodobniej typowe wykonanie pozycji testu z matematyki polega na tym, że uczeń najpierw rozwiązuje zadanie, a potem szuka swojego wyniku wśród opcji odpowiedzi.

[2] Autorzy twierdzą, że zadania testu przypominają problemy, z którymi zmagają się uczeni. Trudno się z tym zgodzić. Uczony nie po to prowadzi badanie, by potem głowić się nad wyjaśnieniem ich wyników.

[3] Po angielsku podobnie: *performance test i test performance*.

(zwykle dopełnienia lub rozstrzygnięcia) zawiera w sobie więcej wskazówek niż typowe wezwanie (por. „Rozważ słowa filozofa «Sam wychowawca musi być wychowany»" i „Czy Juliusz Słowacki wielkim poetą był?"). Innymi słowy — wezwanie wymaga samodzielnej koordynacji czynności wykonawczych.

Badanie trafności testów wykonania (Shavelson i in. 1992) przyniosło rezultaty podobne do poprzednich. Wyniki testów przyrządowych lub wykorzystujących demonstracje przyrodnicze nieźle zgadzają się z wynikami testów wykorzystujących symulacje komputerowe, ale nie zgadzają się z wynikami ani testu krótkich odpowiedzi, ani testu wyboru[4].

Na koniec tego przeglądu wspomnijmy, że wzrasta zainteresowanie modelem SOLO (Biggs i Collis 1982), który mierzy nie tyle dyspozycję, ile strukturę wytworu. W latach 80. i 90. pojawiło się wiele prac na temat budowania pozycje testu i przetwarzania odpowiedzi. Ale entuzjaści SOLO powinni pamiętać, że nie jest to lepsza wersja testu wyboru, lecz test, który mierzy właściwe mu konstrukty.

Z powiedzianego płyną dwa wnioski. Pierwszy — że test wyboru mierzy swoisty zbiór dyspozycji psychicznych i żadną miarą nie może być traktowany jako równoważna wersja testów mających inną formę. Drugi — że wybór formy testu wpływa na alokację przywileju oświatowego, bo wynik testowania wytwarza interakcja cech testu i cech testowanej osoby. Weźmy pouczające badanie Fennema i in. (1998), w którym szukano różnic między dziewczynkami i chłopcami pod względem osiągnięć z matematyki. Autorzy nie wykryli żadnych różnic w poziomie wykonania zadań arytmetycznych, ale wykryli znaczne różnice w sposobie rozwiązywania problemów matematycznych: dziewczynki chętniej stosowały strategie konkretne, np. modelowanie i przeliczanie, a chłopcy — strategie abstrakcyjne, które „odzwierciedlały zrozumienie pojęciowe" (s. 11). Znaczy to, że utrzymaniu przez dziewczynki wysokiej pozycji w szkole sprzyja zastosowanie typowego testu wyboru, a zagraża zastosowanie testu wykonania, testu SOLO itp.

Jeśli rodzaj testu wpływa na pozycję oświatową uczniów, a zatem na dostęp do edukacji dużych grup społecznych, to wybór testu musi być oparty na dobrze uzasadnionym kryterium. Nie jest nim ani taniość, ani wygoda czy nowość. Czy jest takie kryterium? Myślę, że tak, nawet dwa. Kryterium zewnętrzne odwołuje się do pozaoświatowego użytku wyników pomiaru (np. do selekcji personelu). Kryterium wewnętrzne odwołuje się do pojęcia wykształcenia.

Test wyboru a pojęcie wykształcenia

Podzielmy proces kształcenia na sensowne jednostki — takie, że każda ma swoją intencję wyrażoną w języku psychologicznym (np. „wie", „rozumie", „potrafi", „aprobuje lub dezaprobuje"). Test osiągnięć uznamy za dobrze dobrany, jeśli informuje o stopniu, w jakim uczeń zrealizował tę intencję.

Test, jak powiedzieliśmy, to standardowa procedura pobierania próbki interesującego nas zachowania. W doborze testu trzeba wyjść od zainteresowania, a nie od procedury. Jeśli interesuje nas poprawność ortograficzna, wybieramy dyktando. Jeśli interesuje nas umiejętność prowadzenia samochodu, sadzamy ucznia za kierownicą. Jeśli interesują nas umiejętności stawiania diagnozy medycznej, konfrontujemy studenta z pacjentem lub jego symulacją.

Opracowanie procedur pobierania takich próbek i obliczania wyniku jest trudne, a przeprowadzenie testu bywa kosztowne. Nie może to jednak usprawiedliwiać postawy: skoro nie możemy zrobić tego, co trzeba, zróbmy to, co można. Jest ona równie absurdalna, jak postawa wnuczka, który widząc babcię cierpiącą na ból głowy, powiada: trzeba by jej zmierzyć ciśnienie, ale skoro nie mamy ciśnieniomierza, zmierzmy jej temperaturę.

Jeśli ze względów technicznych czy ekonomicznych czynimy podstawą oceny osiągnięć szkolnych test wyboru, bezwiednie zmieniamy pojęcie wykształcenia i zaczynamy z zapałem oferować uczniom coś, czego być może nie uznalibyśmy za wartościowe, gdybyśmy o tym pomyśleli na chłodno. Że taka zmiana rzeczywiście zachodzi, przekonują doniesienia amerykańskiej agencji NAEP *(National Assessment of Educational Progress)*. W 1982 r. aż 90% 17-latków potrafiło wykonać proste dodawanie i odejmowanie, ale wykonanie zadań

[4] Zdaniem autorów wynik testu wykonania bardziej zależy od nauczania, a wynik klasycznych testów — od zdolności ogólnych. Czy tak jest — można wątpić.

wymagających rozumienia twierdzeń matematycznych spadło w ciągu dziesięciolecia z 62% do 58%, a rozwiązywanie problemów z 33% do 29%. Obniżyły się też umiejętności analizowania tekstu, wypowiadania się na piśmie z użyciem umiejętności analitycznych i logicznych oraz rozumowania przyrodoznawczego — czyli wszystko to, czego nie mierzą testy wyboru. Amerykańska szkoła wzięła się za nauczanie umiejętności podstawowych, zaniedbując umiejętności złożone. Jak będzie w Polsce — przekonamy się wkrótce.

Ktoś może zapytać z pozorną naiwnością — dlaczego miałoby to nas martwić? Czyż nie powinniśmy zabiegać o to przede wszystkim, by w ramach publicznego projektu oświatowego wyposażyć całą populację w podstawowe umiejętności? Przeczącą odpowiedź na to ważne pytanie opiera się zwykle na jakimś modelu umysłu. Cytowany Frederiksen (1984) eksponuje rolę konstruktów psychologicznych w rodzaju płynności ideacyjnej i namawia do definiowania wykształcenia za pomocą pojęć psychologii poznawczej. To akurat wydaje mi się wątpliwe: teoria psychologiczna jest zawsze tymczasową rekonstrukcją doświadczenia i nie nadaje się na normatywną podstawę praktyki społecznej. Trzeba szukać innych uzasadnień.

Zauważmy, że testy wyboru (odtwórcze, ustrukturowane i zamknięte) różnią się od innych narzędzi sprawdzania wiedzy głównie tym, że nie wymagają od ucznia samodzielnego koordynowania umysłowych i motorycznych procedur wykonawczych. Gdy uczeń pisze wypracowanie, musi nie tylko przypomnieć sobie fakty i wymyślić argumenty, lecz także uporządkować je wzdłuż jakiejś osi kompozycyjnej, czyli nadać wypracowaniu strukturę. Główne pytanie to: „Od czego zacząć, do czego przejść, na czym skończyć?", a główna troska to: „Czy moja struktura jest niesprzeczna i wyczerpująca?". Takie pytania i troski nie trapią ucznia wykonującego test wyboru, tu bowiem koordynację zapewnia konstruktor. Test jest więc łatwiejszy w tym sensie, że wymaga tylko procedur wykonawczych, a nie kierowniczych, sprawnego wykonywania izolowanych czynności niższego rzędu (wyszukiwania wiadomości, stosowania umiejętności), a nie organizowania pracy (programów umysłowych). Test nie wymaga dystansu wobec reguł (rozluźniania ich w celu znalezienia nowych, bardziej adekwatnych), wysuwania, modyfikowania i porzucania celów, zmieniania priorytetów, tworzenia i korygowania planów.

Na marginesie: czy nie dlatego większość uczniów zdaje się lubić testy? Szkoła nauczyła ich, że edukacja to podążanie nieświadomych za świadomym. Gdy uczą się wraz z rówieśnikami do egzaminu, chcą, by ci zadawali im pytania, a nie by z nimi rozmawiali, przedstawiali własne modele umysłowe, ujawniali luki i niepewność. Test doskonale wpisuje się w takie pojęcie uczenia się. A co z uczniami, którzy nie lubią testów? Czy liczą na swój urok osobisty i inteligencję społeczną, czy może także na swoje zdolności organizowania wytworu tak, by wyeksponować to, co wiedzą, i ukryć to, czego nie są pewni? Czy znajdzie się ktoś, kto podejmie te zagadnienia, zamiast drugorzędnych, często pozornych problemów oceniania osiągnięć szkolnych, od których roją się rozprawy pedagogiczne?

***

Podsumujmy te rozważania. Test wyboru należy do kultury taśmy produkcyjnej. Taśma wypiera rzemiosło, które wymaga, wedle słów Marksa, by wytwór najpierw powstał w głowie wytwórcy. Robotnik ma opanować procedury. O koordynację procedur zadba ktoś inny. Aprobata testu wyboru to aprobata socjalizacji, która wytwarza sprawnych wykonawców. Ale że każde społeczeństwo potrzebuje także koordynatorów — przyjdzie ich wykształcić poza jej strukturami. Socjalizacji sprawnych wykonawców musi więc towarzyszyć socjalizacja sprawnych organizatorów zbiorowego wysiłku.

Czy zatem aprobata testu wyboru nie okazuje się aprobatą oświaty dualnej — podzielonej na szkołę dla mas, które mają nabyć takich „kompetencji", jakie mierzą np. testy Instytutu Badań Kompetencji, i szkołę dla elity, która będzie sterować masami? To nie jest pytanie retoryczne. Stawiam je nie po to, by tanim chwytem zdezawuować wysiłki ludzi, do których należą i organizatorzy tej konferencji, lecz po to, by wypowiedziane, mogło się stać przedmiotem analitycznego i krytycznego dyskursu.

Literatura

Algozzine B. (1991) *Decision making and curriculum-based assessment.* W: B.Y.L. Wong (ed.) *Learning about learning disabilities.* San Diego, Academic Press. Biggs

J.B., Collis K.F. (1982) *Evaluating the quality of learning. The SOLO taxonomy*: New York, Academic Press.

Fennema E., Carpenter T.P., Jacobs V.R., Franke M.L., Levi L.W. (1998) *A longitudinal study of gender differences in young children's mathematical thinking.* „Educational Researcher" 27(5), 6-11.

Frederiksen N. (1984) *The real test bias. Influences of testing on teaching and learning.* „American Psychologist" 39 (3), 193-202.

Konarzewski K. (1985) *Pojęcie niezmiennika w myśleniu dziecka.* „Przegląd Psychologiczny" 28(3), 759-772.

Levine A. G.., McGuire C.H., Nattress L.W. (1970) *The validity of multiple-choice achievement tests as measures of competence in medicine.* „American Educational Research Journal" 7, 69-82.

Rogers E.M. (1986) *Doskonalenie nauczania fizyki poprzez konstrukcje i dyskusje różnych typów sprawdzianów.* Wrocław, Wyd. IKN ODN.

Shavelson R.J., Baxter G.P., Pine J. (1992) *Performance assessments: Political rhetoric and measurement reality*: „Educational Measurement" 21, 22-27.

Simon H.A. (1973) *The structure of ill structured problems.* „Artifical Intelligence" 4, 181-201.